Ensemble Data Assimilation without Ensembles: Methodology and Application to Ocean Data Assimilation

Christian L. Keppenne, Michele M. Rienecker, Robin M. Kovach, and Guillaume Vernieres

Global Modeling and Assimilation Office

Code 610.1
NASA Goddard Space Flight Center
Greenbelt, MD 20771

Abstract

Two methods to estimate background error covariances for data assimilation are introduced. While both share properties with the ensemble Kalman filter (EnKF), they differ from it in that they do not require the integration of multiple model trajectories. Instead, all the necessary covariance information is obtained from a single model integration. The first method is referred-to as SAFE (Space Adaptive Forecast error Estimation) because it estimates error covariances from the spatial distribution of model variables within a single state vector. It can thus be thought of as sampling an ensemble in space. The second method, named FAST (Flow Adaptive error Statistics from a Time series), constructs an ensemble sampled from a moving window along a model trajectory. The underlying assumption in these methods is that forecast errors in data assimilation are primarily phase errors in space and/or time.

Here, SAFE and FAST are applied to the assimilation of Argo temperature profiles into version 4.1 of the Modular Ocean Model (MOM4.1) coupled to the GEOS-5 atmospheric model and to the CICE model developed at Los Alamos National Laboratory. The results are validated against un-assimilated Argo salinity data. They show that SAFE and FAST are competitive with the ensemble optimal interpolation (EnOI) used to produce the latest ocean analysis by the Global Modeling and Assimilation Office (GMAO). Further, when only temperature data are assimilated, FAST is better able to improve the model salinity field than SAFE or EnOI.

1. Introduction

Following a seminal paper by Evensen (1994), the ensemble Kalman filter (EnKF) has become a widely used research tool and research topic. At the time of writing, more than 400 papers have been published on the EnKF and many more on closely related ensemble data assimilation methods. While they differ in terms of the approach used to update or resample the ensemble of model states, all ensemble data assimilation methods require an ad hoc number of concurrent model integrations to estimate the distribution of background errors. This approach (also known as sequential Monte Carlo) is essentially an O($\underline{n}$) procedure, where n is the size of the model state vector. In contrast, the original Kalman (1960) filter algorithm propagates its background-error-covariance estimates by means of matrix multiplications of O($\underline{n}^3$). Hence, ensemble methods are considered economical from a numerical standpoint, even though their cost can be seen as overwhelming in comparison to conventional data assimilation methods. Because of this, nearly all implementations of ensemble assimilation methods must compromise between ensemble size and model resolution.

Because the analysis and error estimates depend on the state of each ensemble member, ensemble methods are considered flow-adaptive. Another attractive property of these methods is that they provide estimates of the cross-field covariance between observed and unobserved model fields that are needed to update unobserved system variables. For example, ocean sub-surface fields can be updated even if only surface observations are available. In what follows we will use the term "multivariate" to refer to the ability to update fields of prognostic model variables other than the observed variable using estimates of the cross-field error covariances. Conversely, algorithms that update only a single observed model field will be referred to as "univariate".

The purpose of this paper is to introduce two data assimilation algorithms that, like the EnKF and other ensemble methods, are flow-adaptive and multivariate but, unlike ensemble methods, rely on only a single model trajectory to estimate all the necessary error-covariance information. As such, these methods obviate the requirement to compromise between ensemble size and model resolution faced by all EnKF implementations for higher resolution numerical models. The first algorithm (Space Adaptive Forecast-error Estimation: SAFE) is based on the concept of sampling an ensemble in space. It estimates error covariances from the joint spatial distribution of model variables in a single background state. The second algorithm (Flow Adaptive error Statistics from a Time series: FAST) estimates covariances from the distribution of an ensemble of high-pass filtered lagged instances of the model state vector sampled along the same trajectory. Because they do not require multiple integrations of the numerical model, SAFE and FAST are overwhelmingly faster and considerably less resource hungry than the EnKF and other ensemble assimilation schemes.

The underlying assumption on which SAFE and FAST are based is that errors in the forecasts used in assimilation are primarily phase errors in space and/or time. For the ocean this assumption makes sense as the dominant source of error can be related to errors in surface forcing, i.e., the timing, intensity, or location of particular atmospheric synoptic events. Thus, the forecast (or background) errors can be related to the timing or intensity in the propagation or advection of oceanic anomalies.

The algorithms are outlined in Section 2 together with the Goddard Earth Observing System (GEOS) modeling and data assimilation system developed at the Global Modeling and Assimilation Office (GMAO). The methodology is compared to conventional assimilation techniques in Section 3 where SAFE and FAST are applied to the assimilation of Argo temperature (T) profiles into the ocean component of the coupled modeling system. Unassimilated Argo salinity (S) observations are used to validate the assimilation. Conclusions follow in Section 4.

2. Assimilation Algorithms

2.1 Preamble

Most sequential data assimilation algorithms are inspired by or derived from the Kalman filter (e.g., Analytical Sciences Corporation Technical Staff 1974) and involve the following steps,

$$\boldsymbol{x}_k^f = \boldsymbol{M}(\boldsymbol{x}_{k-1}^a, \boldsymbol{f}_{k-1}), \quad (1a)$$

$$\boldsymbol{y}_k = \boldsymbol{H}_k(\boldsymbol{x}_k^t) + \boldsymbol{\varepsilon}_k, \qquad E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k^T) = \boldsymbol{R}_k, \quad (1b)$$

$$\boldsymbol{K}_k = \boldsymbol{P}_k^f \boldsymbol{H}_k^T [\boldsymbol{H}_k \boldsymbol{P}_k^f \boldsymbol{H}_k^T + \boldsymbol{R}_k]^{-1}, \quad (1c)$$

$$\boldsymbol{x}_k^a = \boldsymbol{x}_k^f + \boldsymbol{K}_k[\boldsymbol{y}_k - \boldsymbol{H}_k(\boldsymbol{x}_k^f)], \quad (1d)$$

where the subscript *k* refers to the *k*th of a sequence of assimilations, $\boldsymbol{x}^f$ and $\boldsymbol{x}^a$ denote the model forecast and analyzed states, $\boldsymbol{M}$ is the model operator, and $\boldsymbol{f}_{k-1}$ represents the forcing between times t_{k-1} and t_k. The observations, $\boldsymbol{y}_k$, assimilated at time t_k are related to the true system state, $\boldsymbol{x}^t$, at time t_k by equation (1b) where $\boldsymbol{H}_k$ is the observation operator, E denotes the expectation operator and $\boldsymbol{\varepsilon}_k$, with covariance matrix $\boldsymbol{R}_k$, is the observation error vector. The Kalman gain matrix, $\boldsymbol{K}_k$, dictates how the observations and model forecast are weighted in the analysis computation (equation 1d). It depends on $\boldsymbol{H}_k$, $\boldsymbol{R}_k$ and the background error covariance matrix,

$$\boldsymbol{P}_k = E((\boldsymbol{x}^t - \boldsymbol{x}_k^f)(\boldsymbol{x}^t - \boldsymbol{x}_k^f)^T). \quad (2)$$

Of course, since $\boldsymbol{x}^t$ is unknown, $\boldsymbol{P}_k$ cannot be computed directly from equation (2) and must be estimated, either explicitly or implicitly, by some other means. In fact, the procedure used to estimate $\boldsymbol{P}_k$ can be used to classify data assimilation methods.

In the EnKF and most ensemble methods inspired from it, $\boldsymbol{P}_k$ is estimated from the statistical distribution of an ensemble of model forecasts,

$$\boldsymbol{x}_{i,k}^f = \boldsymbol{M}(\boldsymbol{x}_{i,k-1}^a, \boldsymbol{f}_{i,k-1}), \qquad i = 1, \ldots, n, \quad (3)$$

started from an ensemble of n analyzed model states at time t_{k-1}. To filter spurious long-range covariances resulting from the finite ensemble size, nearly all ensemble data assimilation implementations follow Houtekamer and Mitchell (2001) and (dropping the k subscript) decompose $\boldsymbol{P}$ as

$$\boldsymbol{P} = \boldsymbol{P}_e \bullet \boldsymbol{C}, \quad (4)$$

where $\boldsymbol{P}_e$ represents the background covariances estimated from the ensemble of model states, $\boldsymbol{C}$ is a compactly supported correlation matrix and $\bullet$ denotes the Schur (i.e., element by element) product of two matrices.

In a class of methods known alternatively as ensemble optimal interpolation (EnOI: *e.g.,* Borovikov *et al.* 2005; Oke *et al.* 2005, 2010; Wan *et al.* 2010; Vernieres *et al.* 2012) or asymptotic ensemble filters, the time dependency is neglected and $\boldsymbol{P}$ is estimated from the statistics of one or more model run histories or from combinations of model histories. In many cases, EnOI methods are competitive with the EnKF because they make up for the performance degradation due to neglecting the forecast-error evolution by estimating error statistics from a much larger ensemble.

Optimal interpolation (OI: Eliassen 1954) refers to an older class of data assimilation methods in which background-error covariances are modeled with Gaussian functions or other analytically or empirically derived functions. Cross-field covariances are generally neglected in these methods and only the model field corresponding to the observed variable is updated.

2.2 Space Adaptive Forecast error Estimation (SAFE)

The SAFE algorithm attempts to combine the simplicity and cost effectiveness of OI with the large sample size of EnOI and the flow dependency of the EnKF. It estimates background error covariances by sampling each model field in the neighborhood of every grid point and using the resulting sample as if it came from an ensemble of model states. Because the size of the sampling neighborhood determines the covariance amplitudes, rescaling is necessary. Note however that an error-covariance rescaling step is also implicitly present in most ensemble assimilation techniques where the background-error covariance amplitude is determined by parameters of a covariance inflation procedure.

While the concept of sampling errors from an ensemble in space is heuristically convenient, it is impractical in geophysical fluid models with complicated boundaries. Instead, the procedure is approximated with the following algorithm. For simplicity of notation, we assume that the model state can be split according to

$$\boldsymbol{x} = [\boldsymbol{v}, \boldsymbol{w}], \quad \boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}^{vv} & \boldsymbol{P}^{vw} \\ \boldsymbol{P}^{wv} & \boldsymbol{P}^{ww} \end{bmatrix}, \quad (5)$$

where $\boldsymbol{v}$ is an observed model field and $\boldsymbol{w}$ is unobserved. The generalization to more than two model fields is obvious. We also assume that all the data assimilated correspond to the same model field. In view of the above, the model update is split according to

$$\boldsymbol{v}^a = \boldsymbol{v}^f + \underbrace{\boldsymbol{P}^{vv}\boldsymbol{H}^T\left[\boldsymbol{H}\boldsymbol{P}^{vv}\boldsymbol{H}^T + \boldsymbol{R}\right]^{1}\left[\boldsymbol{y} \quad \boldsymbol{H}(\boldsymbol{v}^f)\right].}_{\boldsymbol{v}} \qquad (6a)$$

$$\boldsymbol{w}^a = \boldsymbol{w}^f + \boldsymbol{P}^{wv}\boldsymbol{H}^T\left[\boldsymbol{H}\boldsymbol{P}^{vv}\boldsymbol{H}^T + \boldsymbol{R}\right]^{1}\left[\boldsymbol{y} \quad \boldsymbol{H}(\boldsymbol{v}^f)\right], \qquad (6b)$$

$$= \boldsymbol{w}^f + \boldsymbol{P}^{wv}\left(\boldsymbol{P}^{vv}\right)^{1} \quad \boldsymbol{v}. \qquad (6c)$$

The application of equation (6c) is further simplified by assuming that the *w* background error in grid cell (*i, j, k*) is predominantly related to the *v* error in grid cell (*i, j, k*) and negligibly related to the *v* errors in other grid cells, thus neglecting the off diagonal elements of $\boldsymbol{P}^{vv}$ in (6c). Instead the unobserved model field is updated according to

$$w^a_{ijk} = w^f_{ijk} + \frac{P^{wv}_{ijk}}{P^{vv}_{ijk}} \quad v_{ijk}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad k = 1, \cdots, K, \qquad (6d)$$

where *I*, *J* and *K* denote the number of grid cells along the *x*, *y*, and *z* space dimensions, respectively.

The first step is to estimate the background-error variance of the observed field with

$$\boldsymbol{\sigma}^2_{vv} = diag(\boldsymbol{P}^{vv}) = \Theta([\boldsymbol{v} - \Theta(\boldsymbol{v})]^2), \qquad (7)$$

where Θ is a local 3D averaging operator. Several averaging operators were tested. For our implementation, repetitive application of a gridpoint (spatial) Laplacian smoother was found to be effective.

The variance estimate is rescaled such that

$$\left\| diag(\boldsymbol{H}\boldsymbol{P}^{vv}\boldsymbol{H}^t) \right\| = \gamma^2 \left\| diag(\boldsymbol{R}) \right\|, \quad (8)$$

where the double vertical bar stands for an L2 vector norm. The parameter γ is prescribed. It is a scalar representing the global mean (asymptotic) target ratio of background error variances to data error variances and its role is similar to that of multiplicative covariance inflation parameters used in many EnKF implementations. Note that this formulation assumes a steady state regime where the average global mean error variance increase between successive assimilations equals the mean error variance decrease resulting from each assimilation step.

After estimating the background error variances, the update of equation (6a) is applied. This step corresponds to an OI analysis with the model background error variances calculated with equation (7). Let π_{12} represent the covariance of the *v* background errors at locations 1 and 2. It is estimated with

$$\pi_{12} = \sigma_{v1}\sigma_{v2}\rho_{12}, \quad (9a)$$

$$\rho_{12} = c_0(max(\tfrac{1}{L_v}|v_1 - v_2|, \tfrac{1}{L_x}|x_1 - x_2| + \tfrac{1}{L_y}|y_1 - y_2| + \tfrac{1}{L_z}|z_1 - z_2|)), \quad (9b)$$

where σ_{v1} and σ_{v2} are estimated with equation (7), the *L*s are length scales, in units of the variable *v* and in the three space dimensions and c_0 is the popular function given by equation (4.10) of Gaspari and Cohn (1999), or any other compactly supported correlation function. The *max* function selects the largest of its arguments. Equation (9b) ensures that π_{12} is 0 if either v_1 differs significantly from v_2 or if locations 1 and 2 are very distant from each other. The intent is that in the majority of cases,

$$_{12} = \quad _{v1} \quad _{v2}c_0((|v_1 \quad v_2|)/L_v),$$

and the modulation of the background error covariances with the c_0 function enforces error covariance localization in a state-dependent manner.

The local cross-field covariances of the *v* and *w* errors in every grid cell are estimated with

$$\sigma^2_{vw} = \Theta([\boldsymbol{v} - \Theta(\boldsymbol{v})][\boldsymbol{w} - \Theta(\boldsymbol{w})]). \qquad (10)$$

They are used to update the fields of unobserved variables according to equation (6b).

The size of the regions over which the Θ smoothing operator is applied is generally of little consequence. Figure 1 illustrates this fact. It shows time series of differences in RMS observation minus forecast (OMF) for Argo temperature (T) and salt (S) data when Argo T data are assimilated every five days into the coupled GEOS-5 system, using SAFE. Each panel shows the RMS OMF reduction from the corresponding RMS OMF from a control run without data assimilation, such that negative numbers indicate that the analysis is closer to the data than the control. Fig. 1a corresponds to the active (i.e., assimilated) T data and Fig. 1b to the passive (un-assimilated) S data. As in every other experiment discussed herein, the initial ocean state of each run (including the control run) is taken from the GMAO ocean analysis (Vernieres *et al.* 2012) and the experimental setup is as detailed in Section 3 and uses the model configuration discussed in Section 2.4. In the three cases shown, the Θ operator consists in 5 (red), 10 (blue) and 20 (green) iterations of a diffusive filter. While the case with 20 iterations produces somewhat better results for S (larger RMS S reduction), the differences from the other two cases are small. In the SAFE run of Section 3, Θ involves 10 diffusion steps.

2.3 Flow Adaptive error Statistics from a Time series (FAST)

While SAFE is based on the idea of sampling an ensemble in space, FAST samples an ensemble in time. The analyzed state at time t_k is computed from *n* previous instances of the model state vector sampled from the recent history of the current model run,

$$\boldsymbol{X}_k = \{\boldsymbol{x}_{k-j} - \bar{\boldsymbol{x}}_{(k)}, \quad j = 0, \ldots, n-1\}, \quad (11a)$$

$$\bar{\boldsymbol{x}}_{(k)} = \frac{1}{n}\sum\nolimits_{j=0}^{n-1} \boldsymbol{x}_{k-j}, \quad (11b)$$

where $\boldsymbol{x}_k = \boldsymbol{x}(t_k), \boldsymbol{x}_{k\ 1} = \boldsymbol{x}(t_{k\ 1} = t_k \qquad)$, etc., for a given time lag τ. Arguably, τ should be such that $\boldsymbol{x}_{k-1}$ differs significantly from $\boldsymbol{x}_k$ while it still contains information that is useful at t_k. Hence, it seems reasonable to set L to the assimilation interval, as is done in Section 3.

While one could be tempted to compute the analysis from $\boldsymbol{X}_k$ without further preprocessing as though it were made of the current state of the ensemble of model trajectories from an EnKF run, the resulting error covariance estimates would be dominated by the instances furthest away from the center of the time window since $\bar{\boldsymbol{x}}_{(k)}$ is the simple moving average of length n estimated at time $t_{k-n/2}$. To prevent this from occurring and improve the assimilation performance, the lagged state instances are first high-pass filtered,

$$\boldsymbol{X}'_k = \left\{\boldsymbol{x}_{k-j} - \boldsymbol{x}^0_{k-j}, \quad j = 0,\ldots,n-1\right\}, \quad (12)$$

where the sequence of $\boldsymbol{x}^0_k$ is obtained by low-pass filtering the model history. One simple way to do this is with an exponential moving average (EMA),

$$\boldsymbol{x}^0_k = \alpha\,\boldsymbol{x}_k + (1-\alpha)\,\boldsymbol{x}^0_{k-1}, \quad (13)$$

where $0 \le \alpha \le 1$. A good choice to filter out time scales longer than half the time window is $\alpha = 4/(n+2)$. A simple moving average can be used, as in (11b), at the expense of more bookkeeping. The case with $\alpha = 0.5$ is essentially equivalent to forming the ensemble of first order differences over the time window.

Finally, the $\boldsymbol{X}'_k$ ensemble is resampled. (This step removes sequential ordering information but it is not strictly necessary.) The resulting ensemble mean is removed,

$$\boldsymbol{X}''_k = \left\{\boldsymbol{x}''_{k-j} = \sum\nolimits_{i=0}^{n-1} \beta_{ij}\boldsymbol{x}'_{k-i}, \quad j = 0,\ldots,n-1\right\}, \quad (14a)$$

$$\boldsymbol{X}'''_k = \left\{\boldsymbol{x}''_{k-j} - \bar{\boldsymbol{x}}'', \quad j = 0,\ldots,n-1\right\}, \quad (14b)$$

where the β_{ij} are drawn from a uniform random distribution.

The FAST procedure makes the same calculations to estimate background error covariances and compute assimilation increments with the ensemble of equation (14b) as the EnKF makes with its ensemble of model states at time t_k (*e.g.,* equation 2b-f of Keppenne *et al.* 2008). One notable difference is that FAST calculates only one increment. Because a single model integration is involved, the ensemble size can be increased at a very minimal cost.

Figure 2 illustrates the importance of time-filtering and resampling the lagged ensemble of background states. It shows the reduction in RMS OMF for both T and S with respect to the corresponding RMS OMF statistics from the control run in experiments assimilating Argo T every five days. Five cases are shown. The green lines correspond to the full FAST methodology (equations 11-14) with n=20 and α=0.18 (period 10 EMA). The four other cases shown correspond to (1) the deviations from their ensemble mean of the most recent 20 unfiltered background states sampled every five days (magenta), (2 and 3) the deviations from their ensemble mean of the most recent 20 first order time differences (cyan) and second-order time differences (blue) of background states sampled every five days and (4) the EnOI with the static ensemble of 20 error EOFs used to produce the GMAO analysis (see Section 2.4.2: red). The latter is estimated from a 186-member static ensemble according to the procedure detailed in Vernieres *et al.* (2012).

Clearly, computing covariances from unfiltered background states, a procedure which corresponds to using signal covariances, results in the poorest performance for the passive S data but draws the model state closest to the active T data (possible overfitting). The performance obtained with the dynamic ensembles of most recent first and second order time differences is close to that obtained with the static ensemble of leading EOFs. Among the five cases shown, FAST with 50-day high-pass filtering (period-10 EMA removal from a time series with $L = 5$ days) performs best for salt and achieves a good compromise for temperature. Presumably, the 50-day filtering retains pertinent information and avoids aliasing to the lower frequencies but it is possible that better results could be obtained with a different FAST configuration. The results of Figure 2 are for the entire water column, but it is shown in Section 3 that the better FAST performance is most noticeable above the thermocline.

2.4 GEOS-5 Modeling and Ocean Data Assimilation System

2.4.1 GEOS-5 atmosphere-ocean general circulation model

The SAFE and FAST algorithms are tested in Section 3 in the context of assimilating Argo temperature data into the GFDL MOM4.1 ocean model coupled to the NASA GEOS-5 AGCM and to the Los Alamos CICE ice model (all of which comprise the GEOS-5 AOGCM). The model configuration is the same as that used for the GMAO ocean analysis (Vernieres *et al*. 2012). In summary, the OGCM is run with a geopotential vertical coordinate on a ½° grid with a gradual meridional refinement to ¼° at the Equator and with 40 vertical levels. The grid is Cartesian south of 60°N and tripolar northward thereof. The AGCM grid is 1° × 1.25° with 72 levels. The CICE model is run on the same horizontal grid as the OGCM. The AGCM is constrained by replaying the Modern-Era Retrospective analysis for Research and Applications (MERRA: Rienecker *et al.* 2011) while the ocean observations are assimilated. The replay procedure replaces the AGCM state with the state of the analysis every six hours.

2.4.2 GEOS integrated ocean data assimilation system (iODAS)

The components of the GEOS-5 AOGCM are connected to each other and to the GEOS integrated ocean data assimilation system (iODAS) with the Earth System Modeling Framework (ESMF). Besides SAFE and FAST, two other assimilation algorithms available in iODAS are an EnOI utilizing EOFs of short-term forecast errors (Vernieres *et al.* 2012) and a univariate OI (UOI) algorithm with adaptive error covariance localization. They are both used in Section 3 as comparison benchmarks. The parallel implementation of iODAS follows Keppenne and Rienecker (2003).

The procedure used to compute the short-term forecast-error empirical orthogonal functions (EOFs) used by the EnOI involves a 186-member ensemble of forecast error estimates. It is detailed in Vernieres *et al.* (2012).

The covariance model used by the UOI assumes that the forecast errors have a compactly supported distribution with variance equal to the variance of the ensemble of error estimates from which the EOFs used with EnOI are computed. This corresponds to equation 9 with σ_{v1} and σ_{v2} corresponding to the cumulative EOF standard deviation at locations 1 and 2.

In the SAFE, FAST, EnOI and UOI experiments discussed in Section 3, background-error covariances are localized according to equation (4) where the element of $\boldsymbol{C}$ corresponding to the *i*th and *j*th model state variables at space-time locations (x_i, y_i, z_i, t_i) and (x_j, y_j, z_j, t_j), is given by

$$c_{ij} = c_0(\max(\tfrac{1}{L_r}\left|r_i - r_j\right|, \tfrac{1}{L_x}\left|x_i - x_j\right| + \tfrac{1}{L_y}\left|y_i - y_j\right| + \tfrac{1}{L_z}\left|z_i - z_j\right|))\, c_0(\tfrac{1}{L_t}\left|t_i - t_j\right|), \quad (15)$$

where r_i and r_j are the adaptive localization variable at locations i and j and the r field is chosen to correspond to the observed variable. Note the similarity with equation (9b), except for the appearance of the temporal term, $c_0(\frac{1}{L_t}\left|t_i - t_j\right|)$. The latter results from differences between the measurement times and the analysis time. The application of equation (15) to modulate the background-error covariances enforces a state-dependent error-covariance localization, even when the raw covariances are time-independent, as is the case with EnOI and UOI.

3 Application

To validate the SAFE and FAST algorithms and to evaluate their usefulness as alternatives to the EnOI, we ran four AOGCM experiments assimilating T profiles from the broad-scale global array of temperature/salinity profiling floats (Argo: Gould *et al.* 2004). In three of these runs, namely in the SAFE, FAST and EnOI runs, both T and S ocean model fields are updated. In the fourth run, referred to as UOI run, only the T field is updated. [Note that the UOI is included for completeness, even though it has been known for some time that assimilation that does not update salinity carefully can give a poor analysis (e.g., Sun et al. 2007).] Besides the active Argo T data, passive Argo S data are also processed. The passive qualifier means that these data are not assimilated. Their only use is to quantify the effect of the assimilation on the unassimilated S variable. A control run in which all the T and S data are processed passively was also run.

The runs cover a two-year period starting January 1, 2010. The ocean initial conditions are the same for all runs and come from the GMAO ocean analysis (Vernieres *et al.* 2012). The GEOS-5 replay procedure constrains the atmosphere to MERRA over the period of the runs. Every five days, data from a 5-day time window centered about the analysis time are processed. The observational error model is Gaussian in the horizontal and vertical. The observational error variance varies as a function of depth according to the magnitude of the vertical gradient. Details are provided in Vernieres *et al.* (2012). The assimilation increments are applied incrementally over a five-\day period, as in the incremental analysis update procedure of Bloom *et al.* (1996), but without rewinding the model clock (Keppenne *et al.* 2008).

The SAFE run estimates its background-error covariances from equations (7) and (10) where the Θ operator consists of 10 diffusion steps. To improve the performance in the low latitudes, SAFE error covariances are explicitly disabled when they involve a grid cell within the 10°N-10°S latitude band and another grid cell outside of it. This step is exclusively applied in the SAFE run to prevent the state variables at grid cells outside the waveguide from participating in the estimation of the background error variance (Θ operator) at grid cells inside the waveguide. The FAST run applies equations (11-14) with a five-day lag, n=20 and α=0.18. Only 20 lags are used to facilitate comparison with the EnOI, since the latter uses a static ensemble of 20 leading EOFs (Section 2.4.2). The UOI run takes its background temperature error variance estimate from the EnOI run (weighted sum of squared EOFs). The error-covariance localization scales are the same in all runs and are identical to those used in Vernieres *et al.* (2012). For each assimilation cycle, SSTs are assimilated first, followed by the in situ data.

To illustrate the SAFE and FAST error covariance models, Figure 3 shows time sequences of zonal vertical cross sections at the Equator through the SAFE (Fig. 3 a-d) and FAST (Fig. 3 e-h) background error standard deviation estimates for the model's ocean temperature. The succession is shown with a 3-month interval. The FAST and SAFE sections are qualitatively similar. Yet, the SAFE estimates are noticeably smoother because the number of grid cells participating in the SAFE spatial averaging is larger than the number of lagged state instances used in the FAST computations. Also note the general resemblance to the corresponding section through the time-independent background error standard deviation field used by both the EnOI and UOI runs of

Section 3 (Fig. 3i). The differences between the equatorial sections are largest in the Indian and Atlantic Ocean.

The processing time of each run with data assimilation expressed in terms of the time taken by the control run on 30 Intel Altix Sandy Bridge nodes (360 2.8 GHz cores) is shown in Figure 4. UOI takes 70% longer than the control run while FAST and EnOI both take about twice as long as UOI and SAFE takes nearly 50% longer than UOI. For comparison, the best case scenario for a 20-member EnKF run in which ensemble members are run sequentially is also shown. Running ensemble members in parallel, while possible with the GEOS iODAS would require many more compute nodes.

Figure 5 illustrates the background-error covariance models used in each run by showing marginal T and S assimilation increments corresponding to the impact of a unit T innovation at (0ºN, 140ºW, 180m) at the end of the runs (January 1, 2012). The top row of panels (a), (e) and (i) shows zonal sections through the corresponding marginal T increments in the SAFE (left), FAST (middle) and EnOI (right) runs. The 2nd row of panels (b), (f) and (j) shows corresponding meridional T sections. Panels (c), (g) and (k) (3rd row) and the bottom row of panels (d), (h) and (l) show zonal and meridional sections through the corresponding marginal S increments.

The differences apparent in Figure 5 result primarily from differences in covariance modeling approach (static ensemble in EnOI, time-lagged ensemble in FAST, spatial pseudo ensemble in SAFE). However, differences also arise from differences in the state adaptive error-localization of equation (15) since the differences between the respective background states have increased over time (particularly evident in Figure 6). The amplitude differences between the SAFE, FAST and EnOI marginal gains reflect differences in the background error estimates at the observation location. In this example, there is more correspondence between the shapes of the marginal T and S increments from the EnOI (panels (i) and (k) and panels (j) and (l)) than those from SAFE or FAST. The amplitude of the T marginal increment is also largest in the EnOI run. Yet, the amplitude of the S marginal increment is relatively small in the EnOI run, reflecting lower covariance between the T and S error estimates at this particular observation location.

To further illustrate how the SAFE, FAST and EnOI error-covariance models differ, Figure 6 shows the time evolution (sampled every three months) of zonal sections through the marginal S increment corresponding to a unit T innovation at the same Equatorial location considered in Figure 5. Not surprisingly since the EnOI estimates background covariances from a static ensemble, its marginal S gain at this location displays the least temporal variation. The latter result from how the background T field (r in equation (15)) changes with time. Conversely, the FAST marginal S gain varies the most with time as one could have expected because the corresponding background error covariances are high pass filtered by design and represent errors/uncertainties at periods shorter than 50 days in this case. Clearly, the FAST covariances are influenced by tropical instability waves which mostly occur between July and November and have wavelengths of 1000-2000 km and periods of 20-40 days (e.g., Willett et al., 2006). While the spatial ensemble involved in the SAFE background error covariance calculations also varies with the background fields, the resulting covariances only capture variability in space, not in time.

Figure 7 quantifies the improvement (negative values) or worsening (positive values) over the control by showing to what extent the RMS OMF statistics differ from the corresponding statistics from the control run. RMS OMF differences are shown in each panel for the SAFE (blue), FAST (red), EnOI (green) and UOI (magenta) runs. Figure 7a corresponds to the active Argo T data, while Figs. 7b and 7c correspond to the passive Argo S data above and below 300 meters. While the four data assimilation methods perform similarly for T, FAST stands out for its better performance in terms of S, especially in the upper ocean (Fig. 7b). On the other hand, the underperformance of UOI, which degrades the model salt field compared to the control run, is especially striking in the thermocline (Fig. 7c).

The global RMS observation minus forecast (OMF) differences corresponding to the active T data are comparable in the four runs with T data assimilation (SAFE: 0.76 ºC, FAST: 0.88 ºC, EnOI: 0.76 ºC, UOI: 0.87 ºC), as they each explain approximately the same fraction of the T innovation variance of the control run (1.27^2 $ºC^2$). This result is as expected given that each run sets $\gamma=1$ in equation (8) to facilitate the comparison. Figure 8 further illustrates the respective performance of each run with T assimilation. The difference of the RMS OMF (horizontally and over time) in the data assimilation runs from that in the control is shown as a function of depth for 2011 (blue: SAFE, red: FAST, green: EnOI, magenta: UOI). Negative numbers mean that the data assimilation brings the (5-day lead) forecast state closer to the data than the control and should be the norm if the data are unbiased. Fig. 8a corresponds to the active T data and Fig. 8b to the passive S data. For T, the level of improvement over the control is similar for all runs and is largest near a depth of 100 meters. For S, the results are markedly different. UOI is worse than the control over the entire water column and while SAFE, FAST and EnOI all improve upon the control over the entire column, FAST produces the largest improvement over the entire depth range.

The horizontal distributions of the differences in RMS S OMF from those of the control during 2011 for each of the SAFE, FAST, EnOI and UOI runs are shown in Figure 9 for the upper 300 meters and in Figure 10 for depths greater than 300 meters. In the upper ocean, SAFE, EnOI and UOI all show significant degradations from the control in the Western Equatorial South Pacific (red areas in Figs. 9a, 9c, and 9d). FAST does better in the same area and performs best overall (Fig. 7b). Since the upper ocean salt content is heavily influenced by precipitation and evaporation and the corresponding fluxes are constrained to the MERRA forcing in all runs, including the control, it is not surprising that the analyses (which all assimilate T only) do not outperform the control at the surface and in the mixed layer above the halocline. Positive impacts on the model salinity from the T data assimilation are most likely to manifest themselves further away from the surface. Accordingly, the positive impact of the S field correction in the SAFE, FAST and EnOI runs is more apparent below 300 meters, especially in the Northern Atlantic, Gulf Stream and Kuroshio areas and in the area of the West Australian and Leeuwin currents in the Southeast Indian Ocean. While FAST performs best overall, it under-performs the control in the Indian sector of the Southern Ocean. A similar but less-pronounced underperformance in the same area is also noticeable for SAFE (Fig. 10a), but not for EnOI. Since the comparison is restricted to 2011, these regional comments are not definitive.

4. Conclusions

When ensemble data assimilation schemes are applied to complex numerical models, the ensemble size is always a limiting factor or the object of compromise. The methodologies introduced here are designed to possess the main advantages of ensemble data assimilation methods such as the EnKF, namely the ability to update state variables even if unobserved (or not directly assimilated) and to estimate the spatial distribution of background errors, without incurring the cost of ensemble integrations. In EnKF implementations, the amplitude of ensemble perturbations is often the result of manually tuned covariance inflation. The approach used in SAFE and in FAST whereby the global ratio of the traces of the background error and observation error covariance matrices is specified at runtime achieves essentially the same purpose.

FAST more closely resembles the EnKF since it is essentially an EnKF with a high-pass filtered proxy ensemble sampled from the model's recent history. The spatial background error covariance representation of SAFE is a more radical departure from mainstream ensemble data assimilation methods. While SAFE is nearly as economical as conventional OI, our results hint that it is not as effective as FAST or EnOI in updating fields of un-observed variables. The better performance of FAST in this respect may stem from its error covariance model ability to capture sub-seasonal variability.

While more work is required to fully evaluate the potential of the new methods, our initial results suggest that

SAFE may hold promise for high-resolution data assimilation where numerical efficiency is critical. In complex production systems where running an EnKF implementation requires that the ensemble size or model resolution be severely limited, FAST seems like a viable alternative.

5. Acknowledgement

This work is supported by NASA's Modeling Analysis and Prediction Program under WBS 802678.02.17.01.25. The infrastructure for the runs is provided by the NASA Center for Climate Simulation (NCCS). Yuri Vikhliaev, Max Suarez and Bin Zhao helped configure the GEOS-5 modeling system.

6. Bibliography

Analytical Sciences Corporation Technical Staff, 1974: *Applied Optimal Estimation*, Gelb, A. (Ed.), MIT Press, 374pp.

Bloom, S.C., L.L. Takacs, A.M. DaSilva, and D. Ledvina, 1996: Data assimilation using incremental analysis updates. *Mon. Wea. Rev.*, **124**, 1256-1271.

Borovikov, A., M.M. Rienecker, C.L. Keppenne, and G.C. Johnson, 2005: Multivariate error covariance estimates by Monte-Carlo simulation for assimilation studies in the Pacific Ocean. *Mon. Wea. Rev.*, **133**, 2310-2334.

Eliassen A., 1954: Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. Report 5. Videnskaps Akademiet Institut for Vaer Og Klimaforskning, Oslo, Norway, 12pp.

Evensen, G., 1994: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.*, **99 (C5)**, 10,143-10,162.

Gaspari, G., and S.E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.*, **125B (554)**, 723-757.

Gould, J., D. Roemmich, S. Wijffels, H. Freeland, M. Ignaszewsky, X. Jianping, S. Pouliquen, Y. Desaubies, U. Send, K. Radhakrishnan, K. Takeuchi, K. Kim, M. Danchenkov, P. Sutton, B. King, B. Owens and S. Riser, 2004: Argo Profiling Floats Bring New Era of In Situ Ocean Observations, *EOS, Trans. AGU*, **85 (19)**, 179, 190-191.

Houtekamer, P.L., and H.L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123-137.

Kalman, R., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **D82**, 35-45.

Keppenne, C.L., and M.M. Rienecker, 2003: Assimilation of temperature into an isopycnal ocean general circulation model using a parallel ensemble Kalman filter. *J. Mar. Sys.*, **40-41**, 363-380.

Keppenne, C.L., M.M. Rienecker, J.P. Jacob and R.M. Kovach, 2008: Error covariance modeling in the GMAO ocean ensemble Kalman filter, *Mon. Wea. Rev*., **136**, 2964-2982.

Oke, P.R., A. Schiller, D.A. Griffin and G.B. Brassington, 2005: Ensemble data assimilation for an eddy resolving ocean model, *Q.J. Roy. Met. Soc.*, **131**, 3301-3311.

Oke, P.R.; G.B. Brassington; D.A. Griffin and A. Schiller, 2010: Ocean data assimilation: a case for ensemble optimal interpolation, *Aust. Meteorolog. & Oceanogr. J.*, **59**, 67-76.

Rienecker, M.M., M.J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M.G. Bosilovich, S.D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, et al., 2011. MERRA - NASA's Modern-Era Retrospective Analysis for Research and Applications. *J. Climate*, **24**, 3624-3648.

Vernieres, G., C.L. Keppenne, M.M. Rienecker, J.P. Jacob and R.N. Kovach, 2012: The GEOS-ODAS description and evaluation. Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2012-104606.

Wan, L., L. Bertino and J. Zhu, 2010: Assimilating Altimetry Data into a HYCOM Model of the Pacific: Ensemble Optimal Interpolation versus Ensemble Kalman Filter, *J. Atmos. Ocean. Tech.*, **27 (4)**, 753-765.

Willett, C.S., R.R. Leben, and M. Lavin, 2006: Eddies and tropical instability waves in the eastern tropical Pacific: A review. *Prog. Oceanogr.*, **69**, 218-238.

Figure captions

Figure 1. Reduction of RMS OMF over the corresponding RMS OMF from a control run without data assimilation for (a) active Argo T and (b) passive (i.e., unassimilated) Argo S data in SAFE runs assimilating the Argo T data. The three cases shown correspond to SAFE runs in which the background-error covariance estimation involves 5 (red), 10 (blue) and 20 (green) steps of a diffusive (Laplacian) filter.

Figure 2. Reduction of RMS OMF over the corresponding RMS OMF from a control run without data assimilation for (a) active Argo T and (b) passive (*i.e.,* unassimilated) Argo S data in runs assimilating the Argo T data every five days and in which the background error covariances are estimated with each of the following five approaches (see text for details): EnOI using a static ensemble of 20 leading error EOFs (EnOI: red), a lagged ensemble of the 20 most recent unfiltered background states (0 order: magenta), an ensemble of the 20 most recent first-order time differences (1^{st} order: cyan), an ensemble of the 20 most recent second-order time differences (2^{nd} order: blue), and FAST with 20 lags and 50-day high pass filtering [*i.e.,* removal of a 10-period exponential moving average in equation (12)] (FAST: green). Negative (vs. positive) values correspond to improvements (vs. worsening) over the control.

Figure 3. Temperature background error standard deviation estimates along the Equator in the SAFE, FAST and UOI runs of Section 3 and corresponding from top to bottom to March 31, 2011 (a: SAFE, e: FAST), June 30, 2011 (b: SAFE, f: FAST), September 30, 2011 (c: SAFE, g: FAST) and December 31, 2011 (d: SAFE, h: FAST) Panel (i) shows the time independent background error standard deviation estimate used by both the EnOI and UOI runs. The color scale shown to the right of panel (i) is applicable for all panels.

Figure 4. Processing time per month of model simulation expressed in units of the corresponding processing time from the control run. Note the logarithmic scale. The EnKF case corresponds to a best case scenario for a 20-member EnKF run in which ensemble members are run sequentially.

Figure 5. Zonal and meridional sections through the marginal contribution to the T and S assimilation increments in PSU corresponding to a unit T innovation at (0ºN, 140ºW, 180m) in the SAFE (a-d), FAST (e-h) and EnOI (i-l) runs on January 1, 2012. Zonal (meridional) sections are labeled W-E (S-N). (a), (e), (i) correspond to T zonal sections, (b), (f), (j) to T meridional sections, (c), (g), (k) to S zonal sections and (d), (h), (l) to S meridional sections. The top color bar applies to all the panels in the top two rows. The bottom color bar applies to the bottom two rows.

Figure 6. Zonal sections through the marginal contribution to the S assimilation increment in PSU corresponding to a unit T innovation at (0ºN, 140ºW, 180m) in the SAFE (a-e), FAST (f-j) and EnOI (k-o) runs on (from top to bottom) January 1, 2010, April 1, 2010, July 1, 2010, October 1, 2010 and January 1, 2011. The color bar to the right applies to all the panels.

Figure 7. (a) RMS OMF difference with RMS OMF from a control run without data assimilation started from the same initial condition for (a) active Argo T data, (b) passive Argo S data in the upper 300 meters and (c) passive Argo S data below 3000 meters. RMS OMF differences quantify the improvement (negative values) or worsening (positive values) over the control and are shown in each panel for the SAFE (blue), FAST (red), EnOI (green) and UOI (magenta) runs.

Figure 8. Global average of RMS OMF change over the control as a function of depth for (a) active T data and (b) passive S data in the second year (2011) of the SAFE (blue), FAST (red), EnOI (green) and UOI (magenta) runs. Negative (positive) numbers indicate a reduction (increase) in RMS OMS statistics over the control run.

Figure 9. Horizontal distribution of RMS T OMF differences during 2011 with the corresponding RMS T OMF from the control run. The data are binned over 0-300-meter deep by 1º zonal by 1º meridional boxes. Negative values identify areas where the analysis is closer to the Argo observations than the corresponding state from the control run and vice versa. The four panels correspond to the SAFE (a), FAST (b), EnOI (C) and UOI (d) runs.

Figure 10. Same as Figure 9 for the passive Argo S observations below 300 meters.